PLOS ONE

RESEARCH ARTICLE

Synonymous Co-Variation across the E1/E2 Gene Junction of Hepatitis C Virus Defines Virion Fitness

Brendan A. Palmer*, Liam J. Fanning*

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland

* l.fanning@ucc.ie (LJF); b.palmer@ucc.ie (BAP)

Abstract

Hepatitis C virus is a positive-sense single-stranded RNA virus. The gene junction partitioning the viral glycoproteins E1 and E2 displays concurrent sequence evolution with the 3′-end of E1 highly conserved and the 5′-end of E2 highly heterogeneous. This gene junction is also believed to contain structured RNA elements, with a growing body of evidence suggesting that such structures can act as an additional level of viral replication and transcriptional control. We have previously used ultradeep pyrosequencing to analyze an amplicon library spanning the E1/E2 gene junction from a treatment naïve patient where samples were collected over 10 years of chronic HCV infection. During this timeframe maintenance of an in-frame insertion, recombination and humoral immune targeting of discrete virus sub-populations was reported. In the current study, we present evidence of epistatic evolution across the E1/E2 gene junction and observe the development of co-varying networks of codons set against a background of a complex virome with periodic shifts in population dominance. Overtime, the number of codons actively mutating decreases for all virus groupings. We identify strong synonymous co-variation between codon sites in a group of sequences harbouring a 3 bp in-frame insertion and propose that synonymous mutation acts to stabilize the RNA structural backbone.

Citation: Palmer BA, Fanning LJ (2016) Synonymous Co-Variation across the E1/E2 Gene Junction of Hepatitis C Virus Defines Virion Fitness. PLoS ONE 11(11): e0167089. doi:10.1371/journal.pone.0167089

Editor: Tamir Tuller, Tel Aviv University, ISRAEL

Received: June 29, 2016

Accepted: November 7, 2016

Published: November 23, 2016

Data Availability Statement: The fasta sequence file used to generate the results in this paper has been deposited on figshare (https://figshare.com/articles/RL1-10_fas/4223799).

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Hepatitis C virus (HCV) is genetically diverse. Within the last decade, one new HCV genotype (genotype 7) and 49 new subtypes have been defined [1, 2]. This genetic diversity is founded upon the virally encoded, highly error prone RNA-dependent RNA polymerase [3]. Each HCV genome replication cycle is projected to contain one random mutation event [3, 4]. The accommodation of defined hypervariable regions (HVR) within the HCV genome ensures significant sequence diversity can also be observed at the within-host level. Even at the single nucleotide level, individual sites demonstrating high mutation and low mutation rates have been documented [5].

The potential for exploration of the sequence space is vast, yet collapses in genetic diversity longitudinally have been documented. This may be as a result of the masking of minor variants

by singly dominant sequences, giving the semblance of a clonal phenotype [6]. In one example, analysis of a chronic HCV genotype 4a patient sample yielded just five unique HVR1 amino acid motifs from more than 15,000 individual sequence reads [7]. All five were interconnected by a genetic distance of one amino acid. It is therefore pertinent to examine not only the breadth of the sequence space explored by a virus, but also the apparent limitations behind a lack of exploration [8].

There is increasing evidence that synonymous mutations, which do not alter the amino acid profile, can also have a significant impact on the fitness of viral populations [9, 10]. In a quasispecies of closely related sequences, synonymous codon selection places individual variants in different regions of the sequence space and subsequent changes at these sites can lead to distinct evolutionary trajectories. There is increased recognition that epistasis, the effect(s) of a mutation on the presence or absence of other mutations in the genome, is an important facet of viral fitness. HIV-1 treatment resistance mutations have been observed to occur in tandem with subsequent compensatory mutations to correct for consequential loss of fitness [11]. A number of studies have indicated that significant fitness effects can result solely as a consequence of synonymous mutation [10, 12–14].

In this study, we take advantage of a previously reported data set that is ideally suited to explore epistatic change over an extended period [7]. Firstly, the data spans samples collected over a 10 year period from a HCV chronically infected, treatment naïve patient. Secondly, the sample space was initially comprised of two lineages (L1 and L2) that fluctuated in their dominance of the virome. L1 was of sufficient complexity that sub-lineages L1a, L1b and L1c could be defined. L1 sequences dominated the virome interchangeably for the first 8.6 years of the sampling period (samples 1–8). Each sub-lineage gained temporal dominance within the host and L1a, L1b and L1c were subjected to IgG targeting by the humoral immune response [7]. In addition, towards the latter sampling points, L1 sequence frequency dropped below the levels of detection allowing L2, which was of low genetic diversity and lacked IgG targeting, to rise and dominate the virome *in toto*. Thirdly, L1b sequences contained an in-frame 3 bp insertion within the HVR1. Fourthly, the amplicon spans the E1/E2 gene junction and represents protein elements that are subject to distinct evolutionary pressures. The C-terminal end of E1 forms part of the trans-membrane domain of the protein that engages with capsid protein in the mature virion, whereas the N-terminal end of E2 encodes the antigenic HVR1 [15]. Finally, the sequence set was obtained using ultradeep pyrosequencing. Application of a temporally matched clonal dataset to complement the error correction methodology allowed for considerable sequence depth to be reached [7, 16].

We show disparate patterns of co-variation amongst mutating codon pairs for L1a and L1b sequence subsets. We report that nonsynonymous epistasis dominated the L1a sequence subset and was linked to HVR1 variant change while synonymous co-variation enhanced fitness. We find that synonymous co-variation defined the L1b sequence subset and hypothesize that accommodation of the insertion curtailed mutational flexibility.

Results

Sample set overview and background

The sample set used for this study comprised of ten serum samples from a single, treatment naïve patient, chronically infected with HCV genotype 4a, that were collected over 9.6 years [7]. The mean time between samples was 1.07 years (sd ± 0.43 years). Phylogenetically, the sequence set partitioned into two main lineages named L1 and L2. L1 could be divided further into three sub-lineages, namely L1a, L1b and L1c, based on bootstrap values >85 for each of the main branches [7]. The emergence, dominance and decline of (sub-)lineages are outlined

in S1 Table for reference. L1c sequences were recovered from just 3/10 samples and due to the limited sequence availability were omitted from this analysis. Overall, for L1a and L1b, codon switches were observed to occur at a frequency of 0.071 and 0.079 per site per sequence, respectively. The codon mutation rate for L2 sequences by contrast was 0.028 per site per sequence.

Codon fixation increased over time for all sequence groupings irrespective of sample frequency

Sequences, grouped by (sub-)lineage, were examined at each codon position to identify the mutational flexibility across the length of the amplicon during the study timeframe. L2 contained 42/98 invariant codon sites overall which reflected the low mutation rate of this sequence set. This compares to 16/98 and 19/98 invariant sites within the L1a and L1b sequence subsets, respectively. Unexpectedly, following L2 expansion into the sample space, and despite an increase in the number of unique sequences isolated, the number of actively mutating codon positions decreased over time (Fig 1 and S1 Table).

It was apparent that the accommodation of the in-frame insertion into the HVR1 of L1b sequences negatively impacted on the mutational flexibility of the HVR1. L1b sequences accounted for 96.5% of all sample 7 isolates, yet at that time only four HVR1 codons were actively mutating (and only three thereafter) (Fig 1). This was further evidenced when codon usage frequencies were assessed. Evidence of codon usage bias was present in the data for each of L1a, L1b and L2 (Fig 2). L1b sequence ENC values were consistently lower over the 10 sampling points (minimum ENC 32.4).

Differential patterns of epistasis define (sub-)lineage glycoprotein evolution

Co-variation between codon sites identified epistatic evolution within E1, within E2 and across the E1/E2 gene junction (Fig 3). It was observed that co-variation was deterministic with the same codon-codon switching events being replicated between unique sequences in the majority of instances (S2 and S3 Tables).

Of the 19 co-varying sites that were identified as significant for L1a sequences, 13 formed a highly structured interconnected network. Indeed, all 13 sites had edges connecting to the remaining 12 sites (Fig 3A). This included nonsynonymous-nonsynonymous mutations that crossed the E1/E2 gene junction (codon 349, Figs 3 and 4). Within this network, four sites were synonymous (codon positions 366,367, 377 and 418) with the remaining nine nonsynonymous. Those sites with the greatest significance were nonsynonymous-nonsynonymous interactions within the HVR1 (S4 Table), which was in line with model assumptions.

A more complex and irregular pattern of co-variation was present amongst sites within the L1b sub-lineage (Fig 3B). L1b sites were distributed across the length of the amplicon when compared with the L1a sequence set (Figs 3 and 4). Thirty three sites exhibited significantly linked co-variation to at least one other site. Unlike L1a, the most significant values were not between nonsynonymous-nonsynonymous co-variation within the HVR1, but rather included synonymous change at one or both sites within E1 (Fig 4 and S5 Table). Phylogenetic analysis of just the E1 portion of the amplicon maintained the overall separation of L1a and L1b branches (data not shown). The sequence length was too short (177 bp) for strong phylogenetic inference to be made but this observation indicates that the presence of the insertion applied sequence specific mutational patterns to 3′-end of E1. Furthermore, the majority of L1b sites (19/33) were synonymously mutating throughout.

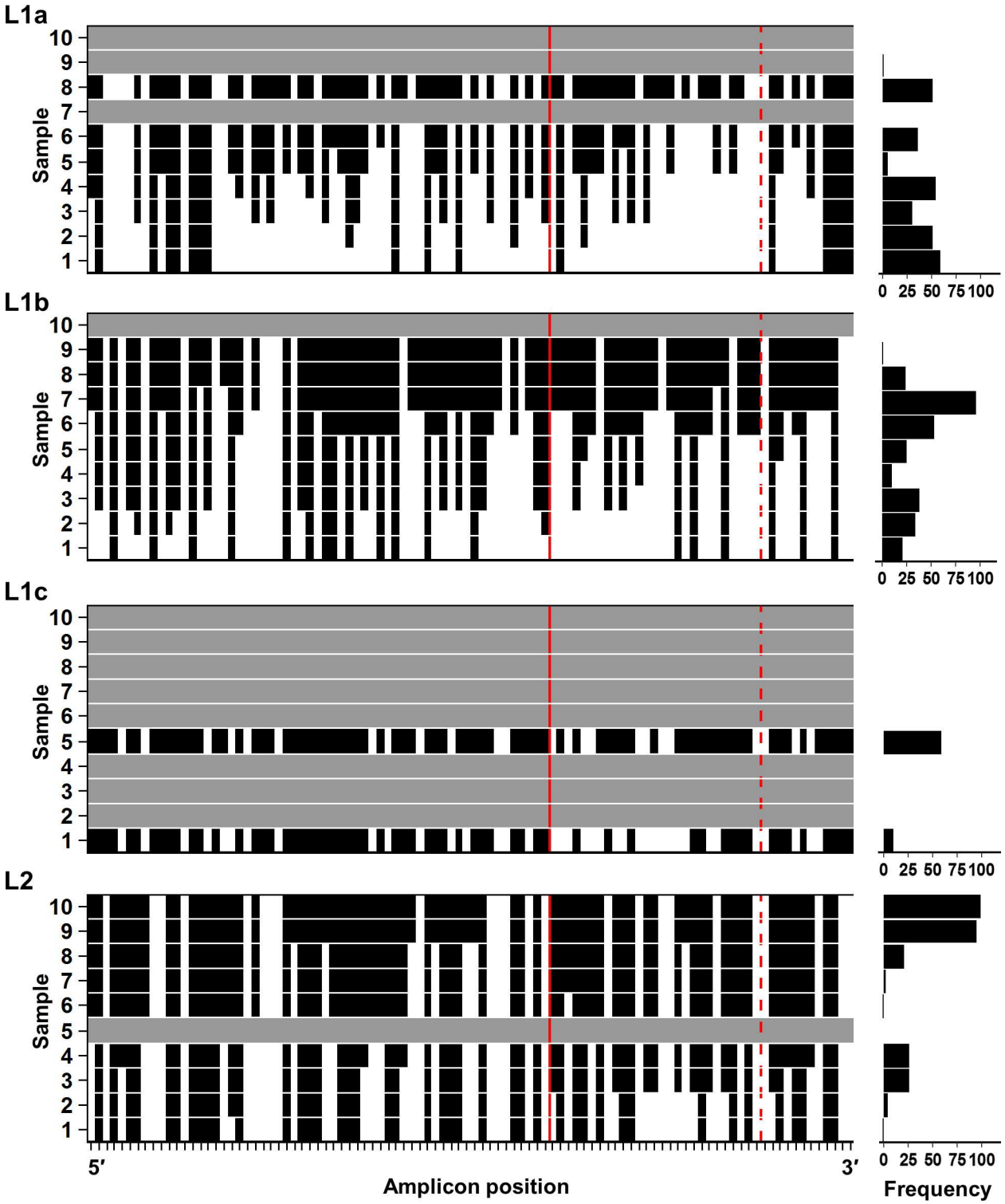


Fig 1. Within-lineage codon fixation during the study timeframe of the L1a, L1b, L1c and L2 sequence subsets. Left panel: For the purposes of this analysis, a site is designated as fixed (black) when a single codon accounts for all the sequences in that sample and all subsequent samples thereafter. For all sequence subsets overtime, the proportion of codon sites actively mutating decreased across the length of the amplicon including the HVR1. Notably, just 3/27 L1b HVR1 codon sites displayed ongoing codon switching events post sample 7. L2 contained the highest proportion of sites that were invariant throughout the sampling timeframe. In spite of the sample space expansion of L2 between samples 8–10, the number of fixed codon sites increased overall. The E1/E2 gene junction and the last codon of the HVR1 are identified by a solid red line and a dashed red line, respectively. Samples with absent or insufficient sequence

data (less than two unique sequences) are shaded as grey horizontal bars. Tick marks along the X-axis identify each codon position of the amplicon sequence. Right panel: The sample specific frequency of each (sub-)lineage.

doi:10.1371/journal.pone.0167089.g001

A single synonymous-synonymous epistatic event was present for L2 sequences at the 3′-end of the amplicon (Fig 3C, codon positions 421 and 422). No significant co-variation was detected in the reference data set using this methodology.

The strength of epistatic interactions contrasted between L1a and L1b sequences. The odds of a codon switching event at a co-evolving site participating in a paired mutational event was high for L1a (Fig 4 and S4 Table). In all instances the combined odds were >0.6. In contrast, a large number of mutations for L1b sequences at significant co-varying sites had an odds ratio <0.6. This observation suggests forced synonymous change to accommodate constraints introduced as a consequence of HVR1 motif drift.

The highly ordered network of epistatic evolution reported for L1a was underscored by the constituent HVR1 amino acid motifs which fell into two discrete groupings that demonstrated

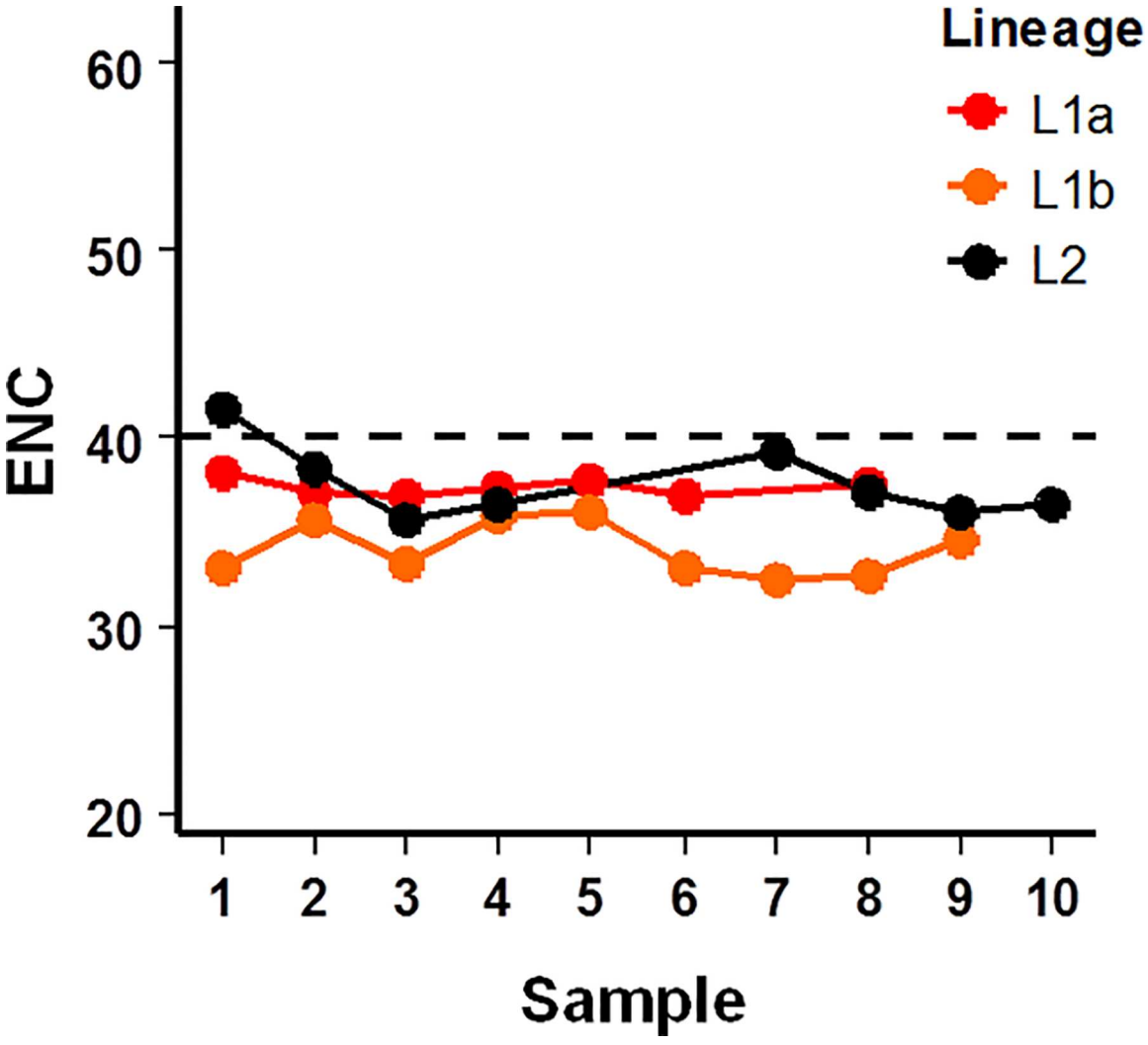


Fig 2. Effective number of codons utilized by each HCV (sub-)lineage overtime. Values less than the threshold of 40 (dashed line) are considered as biased utilization of the available redundancy within the genetic code.

doi:10.1371/journal.pone.0167089.g002

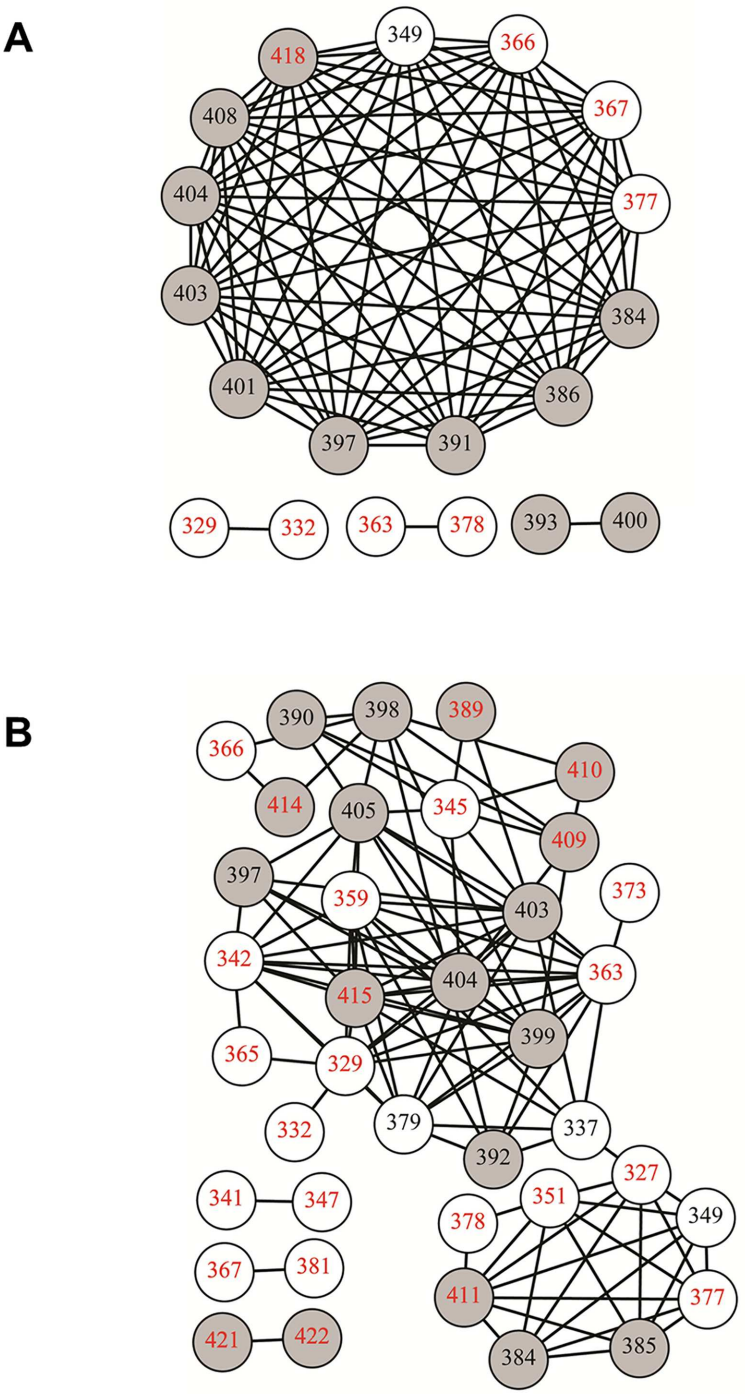

A
349
366
418
367
408
404
377
403
384
401
386
397
391
329
332
363
378
393
400
B
390
398
389
366
410
414
405
345
409
397
359
403
373
342
404
363
415
399
365
329
332
379
337
392
327
341
347
351
349
378
367
381
411
377
421
422
384
385
C
421
422

Fig 3. Pronounced epistasis was evident for both L1a and L1b sequence sets covering 10 years of continuous adaptive evolution of the E1/E2 gene junction. Nodes (representing codon positions) within the graph are connected by an edge if the probability of a change detected simultaneously at both sites was statistically significant (p-value < 0.01). (A) L1a epistasis was highly ordered with the majority of significantly linked sites participating in a single large connected component. (B) Epistasis within the L1b sequence set was observed among a greater number of codon sites overall. The majority of sites identified for L1b exclusively underwent synonymous mutation. (C) Two sites were observed in L2 sequences that were below the significance threshold. White nodes define codons within the E1 coding sequence, while grey nodes identify E2 codons. Sites containing nonsynonymous mutations are identified by black numbers while sites exclusively undergoing synonymous mutation are given by red numbers. Nodes are numbered in accordance with the amino acid positions of the H77 reference genome (Genbank accession: AF009606).

doi:10.1371/journal.pone.0167089.g003

interchangeable dominance sample (S1 Fig). This is in line with the isolation of the same amino acid sequence across multiple samples [17]. L1a HVR1 groups A and B comprised 11 and 8 unique HVR1 amino acid motifs respectively. Within-group evidence of epistasis was limited indicating overall sequence stability (Fig 5).

L1b HVR1 amino acid motifs also partitioned into two groups containing 30 unique motifs (group A) and 9 unique motifs (group B). The closest detectable ancestor to the insertion event, L1a clone GQ985348 isolated from sample 1 [17], branches with L1b group A sequences. Erratic networks of epistatic evolution were detected for this group, which also exhibited the greatest amount of ongoing codon mutation across the length of the amplicon (Fig 4 and S1 Fig). The high proportion of synonymous mutations suggests constraints placed upon the sequence in order to accommodate the insertion initially. Virions harboring the L1b group A motif set never occupied more than 25% of the sample space (S1 Fig). A single codon within the E2 signal peptide (residues 371–383) was identified in nonsynonymous co-varying pairs (Figs 3B and 4, codon 379). However, nonsynonymous codon change accounted for less than 3% of all codon switching events that were recorded at this position and sequences containing the change accounted for just 0.2% of the sample space overall.

The elimination of L1b group A sequences post-sample 5 facilitated the establishment of a fitter, more genetically stable population (group B) that attained near absolute occupation of

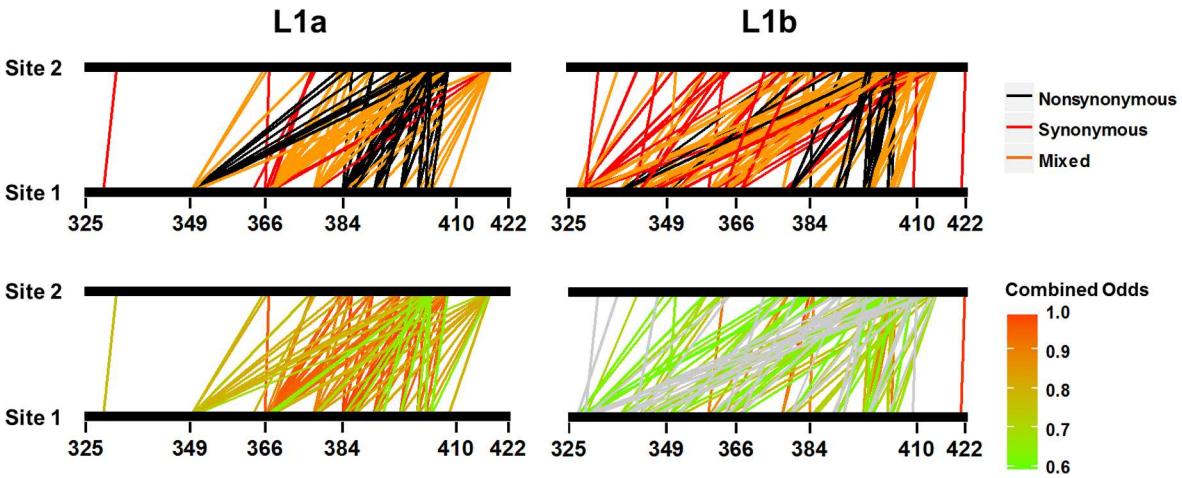


Fig 4. Bipartite mapping of co-evolving sites to the amplicon. Top panel: Co-evolving pairs are identified as nonsynonymous-nonsynonymous, synonymous-synonymous or a combination of one site mutating nonsynonymously and one site mutating synonymously. Bottom panel: The combined odds of a mutation co-varying with a mutation at a second site are given by the colored scale bar. Co-varying pairs represented by grey bars have combined odds <0.6 which indicates that, for a given sequence set, one of the two sites has a greater observed mutational flexibility than that observed at the second site. Raw data counts, individual odds and combined odds are provided in S4 and S5 Tables for reference.

doi:10.1371/journal.pone.0167089.g004

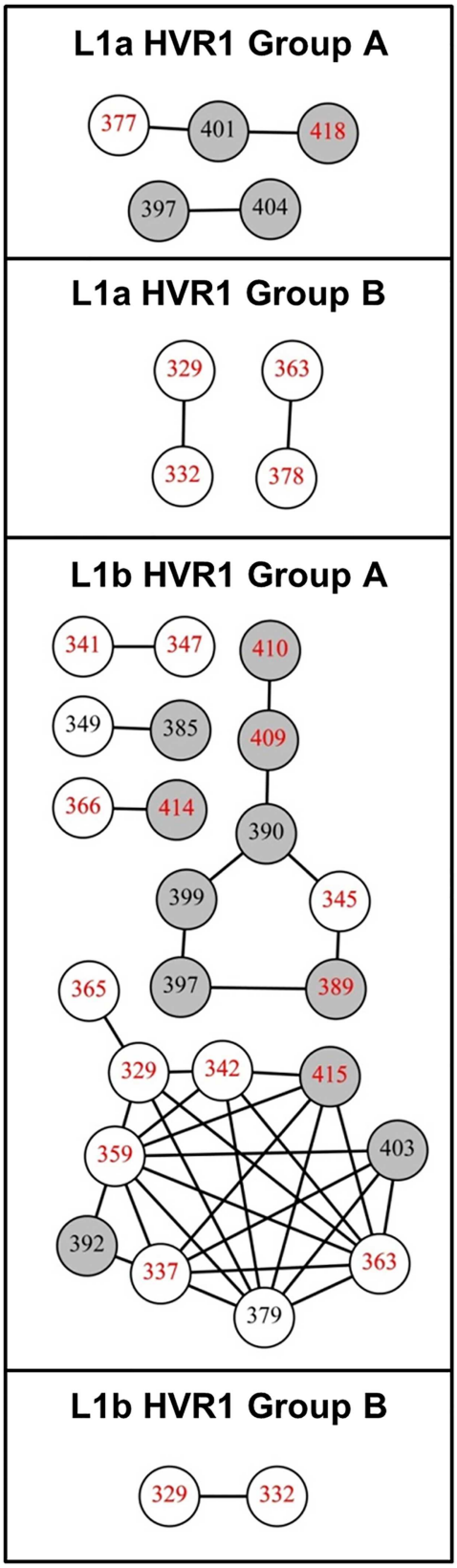

L1a HVR1 Group A
377
401
418
397
404
L1a HVR1 Group B
329
363
332
378
L1b HVR1 Group A
341
347
410
349
385
409
366
414
390
399
345
365
397
389
329
342
415
359
403
392
337
363
379
L1b HVR1 Group B
329
332

Fig 5. Sub-division of L1a and L1b sequences by HVR1 motif defined sequence stability. Significant co-variation between sites was determined separately for L1a sequences and L1b sequences split by HVR1 motif. White nodes define codons within the E1 coding sequence, while grey nodes identify E2 codons. Sites containing nonsynonymous mutations are identified by black numbers while sites exclusively undergoing synonymous mutation are given by red numbers. Nodes are numbered in accordance with the amino acid positions of the H77 reference genome (Genbank accession: AF009606).

doi:10.1371/journal.pone.0167089.g005

the sample space towards the latter end of the sample timeframe (S1 Fig). A single significant synonymous-synonymous co-evolving pair was observed for this sequence subset (Fig 5).

Humoral immune evasion pathways were defined by sequence stability

The separation of L1b sequences by HVR1 motif provided novel insights into the data. L1b group A sequences were phylogentically more diverse whereas group B occupied a narrower sequence space (Fig 6A). Two unique L1b HVR1 amino acid motifs were predicted from the isolation of IgG-bound virions [7]. One of each aligned with L1b group A and group B motifs. Significantly, the L1b group A variant was isolated from sample 1 and represented a small proportion of the overall number of L1b group A sequences (Fig 6B). L1b group A sequences represented 100% of all detectable L1b sequences by sample 5 (Fig 6C and S1 Fig). No evidence of immune targeting was observed for distal branches of L1b group A sequences during this period (Fig 6B).

In spite of this, and compounded by apparent fitness constraints, this subset of the virome collapsed post-sample 5. Interestingly, the L1b group B motif, identified from IgG-bound virions, was not detectable following fractionation of IgG-bound virions until sample 6. It was nevertheless present in unfractionated preparations as early as sample 2. This L1b group B variant was presumably masked from immune targeting by dominant co-circulating variants (Fig 6B and S1 Fig). As L1b group B variants rose to dominate the virome, escape mutations did not develop post-sample 6. All L1b group B variants contained the same HVR1 motif and this motif was associated with IgG-binding (Fig 6B). L1b was at the limit of detection by sample 9 and undetectable by sample 10.

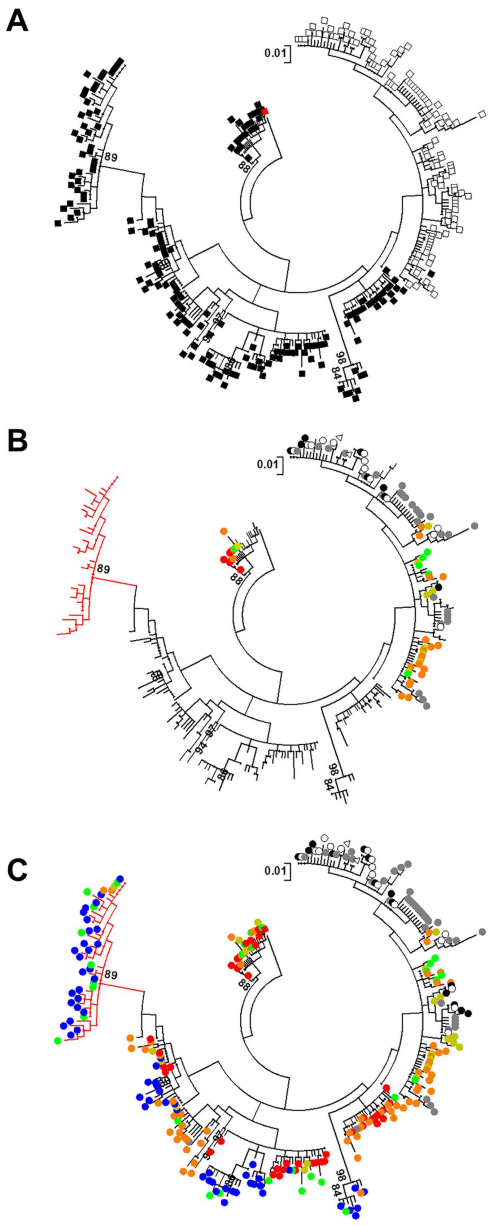
Predicted RNA structures suggest enhanced L1a sequence stability

Discrete RNA structures have been suggested to participate in innate immune evasion and contribute to overall viral fitness [18–20]. Similar secondary structures were predicted within our sequence set. SL1412 as described by Pirakitikulr and colleagues (2016), spans codons 357–371 of the amplicon (Fig 7) [18]. This region is implicated in significant synonymous-synonymous interactions, specifically codons 166 and 167 (Fig 3).

Using the dominant L1a group A and group B sequences as our reference input, the ΔG free energy of the stem loop structure was reduced by the mutations, yet the overall ΔG free energy prediction for the full amplicon was increased. Overall, the impact of individual mutations are unlikely to influence secondary structures which may account for the lack of significant co-variation observed within the L1a group A and B and L1b group B HVR1 sequence subsets. Rather the cumulative effects of mutations across the length of the sequence are the determining factor in genome stability.

Discussion

The use of a well characterized data set spanning 10 years of treatment naïve, chronic HCV infection facilitated the study and comparison of epistatic mutations within co-circulating virus populations. Of the three virus populations described, L1a and L1b exhibited significant



Legend	Sample	Year
●	1	0
●	2	1.8
●	3	2.8
●	4	3.3
●	5	4.6
●	6	6.0
●	7	7.2
○	8	7.6
△	9	8.6

Fig 6. Mapping of the L1b sequence set in the context of HVR1 motif groups. All trees have been rooted using the nearest detectable ancestor to the insertion event, GQ985348 [red square, 17]. The sample specific isolation of each sequence is defined by the color legend. (A) The L1b sequence subset was split into two groups, A (black squares) and B (open squares), which were defined by the constituent HVR1 amino acid motifs. (B) Colored nodes identify those sequences coding for HVR1 motifs that have previously been associated with IgG-bound virions [7]. A single IgG-associated motif from group A was isolated from sample 1 (red circle). An additional IgG-associated motif for group B was isolated from sample 6 (grey circle), four years after the motif was first detected in whole patient serum. A phylogenetically diverse branch of L1b group A sequences not subject to detectable IgG binding is indicated by red branches. (C) Group A sequences not subject to IgG-targeting exhibited the greatest sequence diversity. Nevertheless, this subpopulation collapsed post-sample 5. The genetic distance is shown as a bracketed scale bar. Bootstrap values >80 of 1000 resamplings are shown. Genetic distance is given by the scale bar.

doi:10.1371/journal.pone.0167089.g006

evidence of epistatic evolution. Whereas L1a co-variation was ordered, and both HVR1 subgroups of variants had periods of sample dominance, L1b co-evolution was irregular. The accommodation of an in-frame insertion to the HVR1 of L1b sequences is credited for this difference. L2 sequences had the highest number of invariant sites overall. The L2 sequence subset had the fewest number of unique sequences available for analysis and this in part may account for the higher proportion of invariant sites. It may also reflect an inability to tolerate deleterious mutations, resulting in the efficient removal of low fitness variants from the population [21].

It was apparent at various points of the study, that the insertion event defined the L1b phenotype. The first example was the codon usage bias of L1b sequences. It is recognized that codon usage bias among RNA viruses is largely determined by mutation pressures rather than tRNA availability within the host cell and the viruses own genome nucleotide content [22, 23]. Innate immune recognition of short sequence patterns is also believed to influence codon usage bias in HCV [24]. Consistently, our data shows L1b had lower ENC values than for other (sub-)lineages (Fig 2). The ENC values observed in the current study are considerably lower than those reported by Belalov and Lukashev (2013) who documented a range of 38.2–58.3 for 29 animal RNA viruses [23]. This is likely due to the differing sequence lengths used (full length genomes as opposed to a 294 bp amplicon). Nevertheless, the values for L1b are suggestive of additional constraints placed on L1b sequences.

The ENC results may also have been influenced by the presence of secondary structures in this region of the genome [25–28]. Given that L1b sequences contained an in-frame 3 bp insertion, it is probable that secondary structures in this region would be initially disrupted by the insertion event. For HIV-1 it is well established that gene junctions are enriched in RNA secondary structures and that such structures facilitate recombination-mediated gene swapping [29]. Conserved secondary RNA structures have also been mapped along the entire length of the HCV genome including elements of the E1/E2 gene junction [18–20]. We have previously documented recombinants within our data set and identified the recombination breakpoints at the E1/E2 gene junction [7, 17]. Nascent RNA structures have been known to stimulate transcriptional pausing or transcript release for some time [30]. Synonymous mutations have the potential to discretely alter these structures influencing the outcome of genomic replication events. Identification of a synonymous mutation stably maintained during infection was demonstrated to increase the stability of the RNA structures [19, 31]. Local RNA structures form part of genome-scaled ordered RNA structures (GORS) and such genome organization can be found in many mammalian RNA viruses [28, 32]. GORS form networks of regulatory structures and an understanding of their roles is the focus of ongoing research [18, 19, 28, 33]. While the cumulative impact to the virion is difficult to define, the analysis of localized structures has identified a number of roles including innate immune masking, fitness, replication and infectivity effects [18, 19, 34].

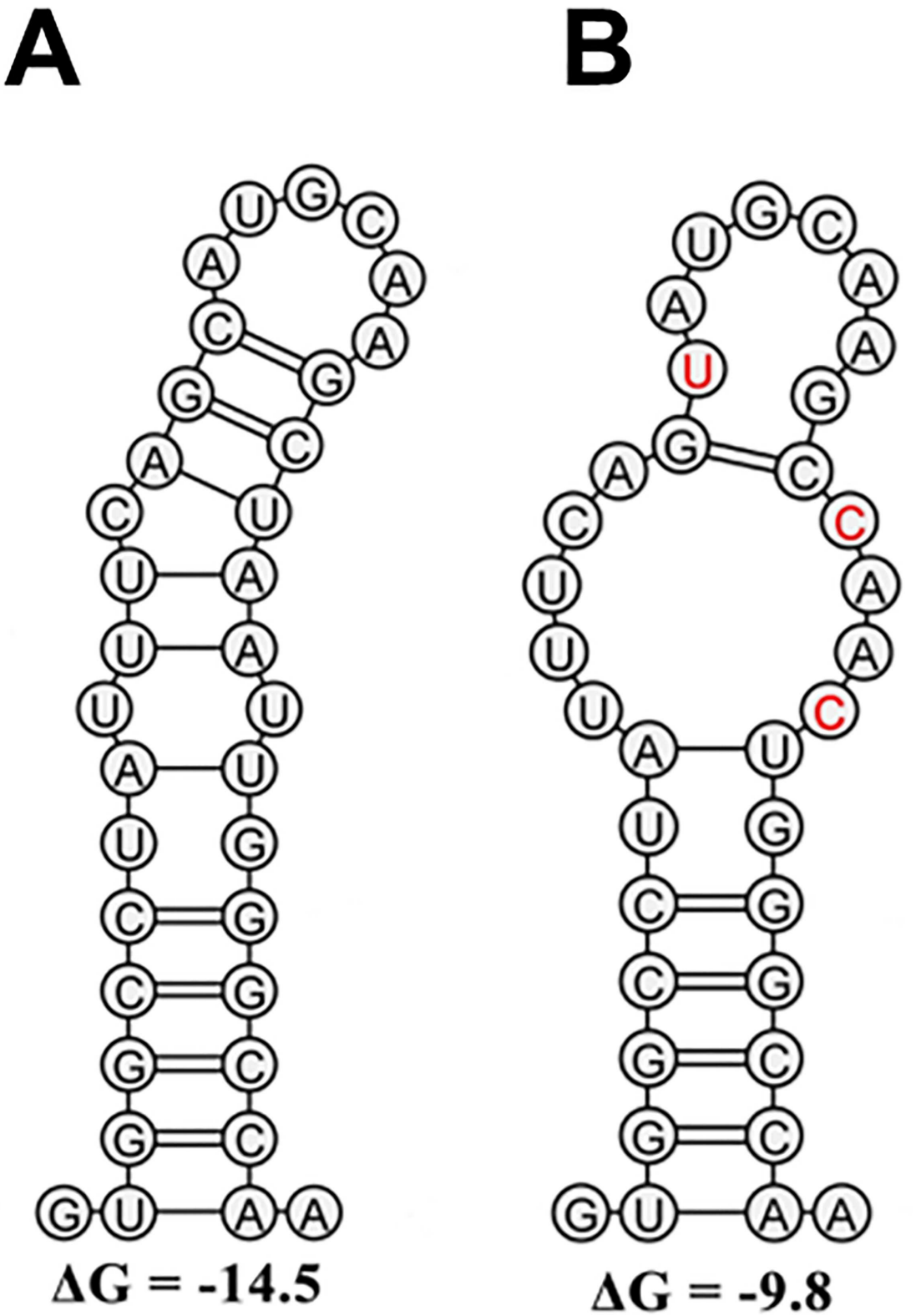


Fig 7. Putative reorganization of conserved structures within the HCV genome over time. The local RNA structure for codons 357–370 were modeled using example sequences from L1a. (A) L1a dominant sequence motif from sample 1. (B) L1a dominant sequence motif from sample 8. Whereas the mutations observed led to an overall decrease in the minimum free energy of the structure, it is the cumulative contribution of mutations across the length of the amplicon that

determine overall stability. The predicted ΔG for the full length amplicon was -93.94 and -100.4 for the dominant sequence motifs from sample 1 and 8, respectively. Synonymous mutations at codons 363, 366 and 367 are shown in red.

doi:10.1371/journal.pone.0167089.g007

Specifically, mutations that strengthened and mutations that weakened a stem loop region of the HCV genotype 2 Jc1 infectious clone, encompassing codons 357–371 of the amplicon reported here, significantly diminished and improved infectivity, respectively [18, 35]. This region formed part of the synonymous component of the L1a epistatic network (Fig 3). Additionally, the synonymous-synonymous pairing of codons 366 and 367 was the most significant of all synonymous-synonymous recorded for L1a. It is probable that the pairwise change was directly linked to the predicted stem loop structure of this region (Fig 7).

We observed significant nonsynonymous co-variation across the E1/E2 gene junction at discrete sites (most notably codon 349 for both L1a and L1b). Extensive, genome-wide networks of HCV protein co-evolution have been described which suggest intra-cellular interactions [36, 37]. Weak evidence of predicted co-variation between the E1 trans-membrane domain and E2 HVR1 has previously been presented [36]. Together, the data suggests intra-cellular processes between viral proteins will result in the development of vast co-variation networks. Ongoing evolution over extended periods in a single host may allow for the linkage of discrete site to these networks additively, enhancing overall fitness.

Whether the net effect of epistasis observed in this study was antagonistic or synergistic can be evidenced from the relative dominance of variant groups. The deterministic facet of co-variation within the L1a sequence set was nonsynonymous mutations at the HVR1 through required preservation of the regions' physio-chemical properties [38, 39]. The consequent genetic drift was accompanied by periodic dominance of the newly emergent variants (S1 Fig). Here epistasis was contributing to adaptation and, cumulatively, new variants emerged to dominate. Conversely, the prevalence of synonymous sites participating in co-variation with nonsynonymous change at the HVR1 suggests compensatory evolution at secondary sites within the L1b group A sequence set (Figs 3 and 4). This was compounded by the disproportionate odds of a mutation participating in a co-evolutionary event at one site over the other (Fig 5 and S5 Table). The co-variation profile of L1b demonstrated that group A variants were unfit and such mutations were likely antagonistic. L1b group B variants were fit, yet could not overcome eventual IgG-targeting through HVR1 genetic drift (Fig 6 and S1 Fig).

The perception that synonymous change equates to 'silent' change is false [9, 10, 12, 40]. In this report, we add to observations that synonymous substitutions occurring amongst related viral variants may hold answers with respect to viral fitness and evolutionary strategies [9, 14, 41]. Building on previous definitions of virus lineage, dominance removal and emergence, we have demonstrated clear evidence of directed synonymous change across the E1/E2 gene junction of HCV [7, 17]. Increasingly, longitudinal reports of HCV infection over decades have identified collapses in sequence heterogeneity, yet viral infection persists. Our data indicates that the acquisition and accommodation of host adaptations over prolonged periods is, in part, maintained and governed at the level of synonymous mutation.

Materials and Methods

Sample data set and preparation

The primary data set on which this study was based has been previously reported [7]. The fasta sequence file used to generate the results in this paper has been deposited on figshare (https://figshare.com/articles/RL1-10_fas/4223799). Those sequences above a sample specific

frequency cut off of 0.1% were retained for downstream analysis and known recombinant sequences were removed [7]. Primer sites at the 5′- and 3′-ends of all sequences were clipped resulting in a final fragment of length 294 bp available for analysis, corresponding to positions 1311–1604 of reference genotype 1a strain H77 (Genbank accession: AF009606). The sequence data set was analysed by (sub-)lineage.

Reference data

A total of 50 unrelated complete genotype 4 genome sequences were retrieved from the Los Alamos National Laboratory HCV sequence database [42]. All sequences were aligned and trimmed to match the 294 bp (297 bp for L1b) E1/E2 region using MEGA6 [43].

Identification of epistatic linkages

Sequences were grouped based on phylogenetic analysis into four sets, namely L1a, L1b, L1c and L2 [7]. Initially the most frequently occurring codon by site (from nonredundant sequence data across all ten samples) was identified and set as the baseline from which change to an alternate codon could be identified. Within each sequence where codon change occurred, simultaneous site changes with other within-sequence sites were identified and the overall inter-sequence observances of matches enumerated.

The number of simultaneous changes across all sequences is considered to have a binomial distribution where 1 applies to a change occurring at both site A and site B simultaneously, and 0 where there is change at site A but not site B, there is change at site B but not site A or there is no change at site A and site B. The observed probability of a change occurring at each individual site (i.e. P(A), P(B), P(C), P(D). . .. . .) was calculated by enumerating the number of observed changes at that site and dividing by the number of potential changes (the number of unique sequences present). In the absence of co-evolutionary constraints, change at individual sites was hypothesised to occur independently of change at subsequent sites. Therefore, the probability of change occurring simultaneously at site A and site B is the product of the individual probabilities of a change occurring at site A and site B. The binomial parameter, p, is estimated to be this product of the probability of a change occurring at site A and the probability of change occurring at site B and the binomial parameter n is the number of sequences available for analysis.

The observed probability of a change occurring simultaneously at any two sites, p_{obs}, was calculated by enumerating the number of observed simultaneous changes at that pair of sites and dividing by the number of potential changes (the number of unique sequences present). The probability of simultaneous site change occurring independently, p, was then compared against the observed probability of simultaneous site change, p_{obs} using the binomial test. This test was carried out for all site pairs in the sequence. An important caveat to the assumption of independent between-site co-variation is that mutational change at the HVR1 is known to require physio-chemical conservation of the region [38, 39]. This informed post-hoc analysis of the data.

The resultant p-value from the binomial test of mutual change occurring at two sites was adjusted to account for Type I errors generated by multiple comparisons using the False Discovery Rate (FDR) procedure [44]. For the purposes of our analysis, all FDR-adjusted p-values <0.01 were deemed significant. Furthermore, the strength of any co-variation was ordered by the associated FDR-adjusted p-value (S4 and S5 Tables). Analyses were performed using R version 3.1.3 [45].

Downstream analysis of co-varying pairs identified by the above procedure is presented in terms of sites mutating nonsynonymously and synonymously. Nonsynonymous sites are defined as an amino acid change occurring in one or more sequences over all sequences

analyzed. Synonymous sites have undergone synonymous change exclusively over all sequences analyzed.

Bioinformatic analyses

Phylogenetic analysis of the L1b sub-lineage was performed using MEGA6 [43]. First, the sequence data was analysed using jModelTest to determine an appropriate model of nucleotide substitution [46]. A general time-reversible model was chosen with gamma-distributed and invariant sites (GTR+G+I). Bootstrap resampling (1000 datasets) of the multiple alignments was used to test the statistical robustness of the trees.

The effective number of codons (ENC) was enumerated using the "chips" module hosted by EMBOSS [47]. ENC is an intuitive measure of codon usage bias. Values range from a low of 20 (only one codon per amino acid used) to a high of 61 (i.e. the use of alternative synonymous codons is equally likely).

RNA secondary structures and free energies were predicted using mfold [48]. Structures were rendered using VARNA [49].

Supporting Information

S1 Fig. Within-sublineage codon fixation during the study timeframe. L1a and L1b sequence sets were split into two groups based on HVR1 amino acid motif. (A) Left panel: For the purposes of this analysis, a site is designated as fixed (black) when a single codon accounts for all the sequences in that sample and all subsequent samples thereafter. Samples with absent or insufficient sequence data are shaded grey. The E1/E2 gene junction and the last codon of the HVR1 are identified by a solid red line and a dashed red line, respectively. Samples with absent or insufficient sequence data (less than two unique sequences) are shaded as grey horizontal bars. Tick marks along the X-axis identify each codon position of the amplicon sequence. Right panel: The sample specific frequency of each (sub-)lineage. (B) Consensus HVR1 motif sequences for L1a, groups A and B, and L1b, groups A and B.
(TIF)

S1 Table. Initial sequence numbers and lineage frequencies
(XLSX)

S2 Table. Frequency of codon changes among co-evolutionary sites observed in L1a sequences
(XLSX)

S3 Table. Frequency of codon changes among co-evolutionary sites observed in L1b sequences
(XLSX)

S4 Table. Odds of participation in a significant co-evolutionary events in L1a sequences ranked in order of p-value
(XLSX)

S5 Table. Odds of participation in a significant co-evolutionary events in L1b sequences ranked in order of p-value
(XLSX)

Author Contributions

Conceptualization: BAP LJF.

Data curation: BAP.

Formal analysis: BAP.

Investigation: BAP.

Methodology: BAP LJF.

Resources: LJF.

Software: BAP.

Supervision: LJF.

Validation: BAP.

Visualization: BAP.

Writing – original draft: BAP LJF.

Writing – review & editing: BAP LJF.

References

1. Smith DB, Bukh J, Kuiken C, Muerhoff AS, Rice CM, Stapleton JT, et al. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. Hepatology. 2014; 59(1):318–27. Epub 2013/10/12. doi: 10.1002/hep.26744 PMID: 24115039; PubMed Central PMCID: PMC4063340.
2. Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, et al. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology. 2005; 42(4):962–73. Epub 2005/09/09. doi: 10.1002/hep.20819 PMID: 16149085.
3. Steinhauer DA, Domingo E, Holland JJ. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene. 1992; 122(2):281–8. PMID: 1336756.
4. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008; 9(4):267–76. Epub 2008/03/06. nrg2323 [pii] doi: 10.1038/nrg2323 PMID: 18319742.
5. Geller R, Estada U, Peris JB, Andreu I, Bou JV, Garijo R, et al. Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat Microbiol. 2016; 1(7):16045. doi: 10.1038/nmicrobiol.2016.45 PMID: 27572964.
6. Palmer BA, Schmidt-Martin D, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, et al. Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses. J Virol. 2015; 90(7):3318–29. doi: 10.1128/JVI.02995-15 PMID: 26719263; PubMed Central PMCID: PMCPMC4794698.
7. Palmer BA, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis C virus mapped by ultradeep pyrosequencing. J Virol. 2014; 88(23):13709–21. Epub 2014/09/19. doi: 10.1128/JVI.01732-14 PMID: 25231312; PubMed Central PMCID: PMC4248971.
8. Holmes EC. Error thresholds and the constraints to RNA virus evolution. Trends in microbiology. 2003; 11(12):543–6. PMID: 14659685.
9. Lauring AS, Acevedo A, Cooper SB, Andino R. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe. 2012; 12(5):623–32. doi: 10.1016/j.chom.2012.10.008 PMID: 23159052; PubMed Central PMCID: PMC3504468.
10. Cuevas JM, Domingo-Calap P, Sanjuan R. The fitness effects of synonymous mutations in DNA and RNA viruses. Mol Biol Evol. 2012; 29(1):17–20. doi: 10.1093/molbev/msr179 PMID: 21771719.
11. Handel A, Regoes RR, Antia R. The role of compensatory mutations in the emergence of drug resistance. PLoS computational biology. 2006; 2(10):e137. doi: 10.1371/journal.pcbi.0020137 PMID: 17040124; PubMed Central PMCID: PMCPMC1599768.
12. Hillung J, Cuevas JM, Elena SF. Evaluating the within-host fitness effects of mutations fixed during virus adaptation to different ecotypes of a new host. Philos Trans R Soc Lond B Biol Sci. 2015; 370 (1675). doi: 10.1098/rstb.2014.0292 PMID: 26150658; PubMed Central PMCID: PMCPMC4528490.

13. Sanjuan R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A. 2004; 101(22):8396–401. doi: 10.1073/pnas.0400146101 PMID: 15159545; PubMed Central PMCID: PMCPMC420405.

14. Kawaguchi K, Faulk K, Purcell RH, Emerson SU. Reproduction in vitro of a quasispecies from a hepatitis C virus-infected patient and determination of factors that influence selection of a dominant species. J Virol. 2011; 85(7):3408–14. doi: 10.1128/JVI.02554-10 PMID: 21270157; PubMed Central PMCID: PMCPMC3067868.

15. Dubuisson J. Hepatitis C virus proteins. World J Gastroenterol. 2007; 13(17):2406–15. PMID: 17552023. doi: 10.3748/wjg.v13.i17.2406

16. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, et al. Efficient error correction for next-generation sequencing of viral amplicons. BMC bioinformatics. 2012; 13 Suppl 10:S6. Epub 2012/07/13. doi: 10.1186/1471-2105-13-S10-S6 PMID: 22759430; PubMed Central PMCID: PMC3382444.

17. Palmer BA, Moreau I, Levis J, Harty C, Crosbie O, Kenny-Walsh E, et al. Insertion and recombination events at hypervariable region 1 over 9.6 years of hepatitis C virus chronic infection. J Gen Virol. 2012; 93(Pt 12):2614–24. Epub 2012/09/14. doi: 10.1099/vir.0.045344–0 PMID: 22971825.

18. Pirakitikulr N, Kohlway A, Lindenbach BD, Pyle AM. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. Mol Cell. 2016; 62(1):111–20. doi: 10.1016/j.molcel.2016.01.024 PMID: 26924328; PubMed Central PMCID: PMCPMC4826301.

19. Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, et al. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc Natl Acad Sci U S A. 2015; 112(12):3692–7. doi: 10.1073/pnas.1416266112 PMID: 25775547; PubMed Central PMCID: PMCPMC4378395.

20. Stewart H, Bingham RJ, White SJ, Dykeman EC, Zothner C, Tuplin AK, et al. Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging. Scientific reports. 2016; 6:22952. doi: 10.1038/srep22952 PMID: 26972799; PubMed Central PMCID: PMCPMC4789732.

21. Elena SF, Sole RV, Sardanyes J. Simple genomes, complex interactions: epistasis in RNA virus. Chaos. 2010; 20(2):026106. doi: 10.1063/1.3449300 PMID: 20590335.

22. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 2003; 92(1):1–7. PMID: 12606071.

23. Belalov IS, Lukashev AN. Causes and implications of codon usage bias in RNA viruses. PLoS One. 2013; 8(2):e56642. doi: 10.1371/journal.pone.0056642 PMID: 23451064; PubMed Central PMCID: PMCPMC3581513.

24. Washenberger CL, Han JQ, Kechris KJ, Jha BK, Silverman RH, Barton DJ. Hepatitis C virus RNA: dinucleotide frequencies and cleavage by RNase L. Virus Res. 2007; 130(1–2):85–95. doi: 10.1016/j.virusres.2007.05.020 PMID: 17604869; PubMed Central PMCID: PMCPMC2186174.

25. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009; 324(5924):255–8. doi: 10.1126/science.1170160 PMID: 19359587; PubMed Central PMCID: PMC3902468.

26. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010; 107(8):3645–50. doi: 10.1073/pnas.0909910107 PMID: 20133581; PubMed Central PMCID: PMC2840511.

27. Shah P, Gilchrist MA. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. PLoS Genet. 2010; 6(9):e1001128. doi: 10.1371/journal.pgen.1001128 PMID: 20862306; PubMed Central PMCID: PMC2940732.

28. Simmonds P, Tuplin A, Evans DJ. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA. 2004; 10(9):1337–51. Epub 2004/07/27. doi: 10.1261/rna.7640104 PMID: 15273323; PubMed Central PMCID: PMC1370621.

29. Simon-Loriere E, Martin DP, Weeks KM, Negroni M. RNA structures facilitate recombination-mediated gene swapping in HIV-1. J Virol. 2010; 84(24):12675–82. Epub 2010/10/01. JVI.01302-10 [pii] doi: 10.1128/JVI.01302-10 PMID: 20881047; PubMed Central PMCID: PMC3004330.

30. Artsimovitch I, Landick R. Interaction of a nascent RNA structure with RNA polymerase is required for hairpin-dependent transcriptional pausing but not for transcript release. Genes & development. 1998; 12(19):3110–22. Epub 1998/10/09. PMID: 9765211; PubMed Central PMCID: PMC317188.

31. Yi M, Hu F, Joyce M, Saxena V, Welsch C, Chavez D, et al. Evolution of a cell culture-derived genotype 1a hepatitis C virus (H77S.2) during persistent infection with chronic hepatitis in a chimpanzee. J Virol. 2014; 88(7):3678–94. doi: 10.1128/JVI.03540-13 PMID: 24429362; PubMed Central PMCID: PMCPMC3993530.

32. Davis M, Sagan SM, Pezacki JP, Evans DJ, Simmonds P. Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. J Virol. 2008; 82(23):11824–36. doi: 10.1128/JVI.01078-08 PMID: 18799591; PubMed Central PMCID: PMCPMC2583674.

33. Fricke M, Marz M. Prediction of conserved long-range RNA-RNA interactions in full viral genomes. Bioinformatics. 2016. doi: 10.1093/bioinformatics/btw323 PMID: 27288498.

34. Witteveldt J, Blundell R, Maarleveld JJ, McFadden N, Evans DJ, Simmonds P. The influence of viral RNA secondary structure on interactions with innate host cell defences. Nucleic Acids Res. 2014; 42 (5):3314–29. doi: 10.1093/nar/gkt1291 PMID: 24335283; PubMed Central PMCID: PMCPMC3950689.

35. Pietschmann T, Kaul A, Koutsoudakis G, Shavinskaya A, Kallis S, Steinmann E, et al. Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. Proc Natl Acad Sci U S A. 2006; 103(19):7408–13. doi: 10.1073/pnas.0504877103 PMID: 16651538; PubMed Central PMCID: PMCPMC1455439.

36. Champeimont R, Laine E, Hu SW, Penin F, Carbone A. Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. Scientific reports. 2016; 6:26401. doi: 10.1038/srep26401 PMID: 27198619; PubMed Central PMCID: PMCPMC4873791.

37. Hagen N, Bayer K, Rosch K, Schindler M. The intraviral protein interaction network of hepatitis C virus. Mol Cell Proteomics. 2014; 13(7):1676–89. doi: 10.1074/mcp.M113.036301 PMID: 24797426; PubMed Central PMCID: PMCPMC4083108.

38. Penin F, Combet C, Germanidis G, Frainais PO, Deleage G, Pawlotsky JM. Conservation of the conformation and positive charges of hepatitis C virus E2 envelope glycoprotein hypervariable region 1 points to a role in cell attachment. J Virol. 2001; 75(12):5703–10. PMID: 11356980. doi: 10.1128/JVI.75.12.5703-5710.2001

39. Hino K, Korenaga M, Orito E, Katoh Y, Yamaguchi Y, Ren F, et al. Constrained genomic and conformational variability of the hypervariable region 1 of hepatitis C virus in chronically infected patients. J Viral Hepat. 2002; 9(3):194–201. Epub 2002/05/16. PMID: 12010507.

40. Sanjuan R, Moya A, Elena SF. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci U S A. 2004; 101(43):15376–9. Epub 2004/10/20. doi: 10.1073/pnas.0404125101 PMID: 15492220; PubMed Central PMCID: PMC524436.

41. Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol Mol Biol R. 2012; 76(2):159–216. doi: 10.1128/Mmbr.05023-11 PMID: WOS:000305508000002.

42. Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics. 2005; 21(3):379–84. Epub 2004/09/21. doi: 10.1093/bioinformatics/bth485 PMID: 15377502.

43. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013; 30(12):2725–9. doi: 10.1093/molbev/mst197 PMID: 24132122; PubMed Central PMCID: PMC3840312.

44. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001; 125(1–2):279–84. PMID: 11682119.

45. R Core Team (2015). R: A language and environment for statistical computing. URL https://www.R-project.org/.

46. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008; 25(7):1253–6. Epub 2008/04/10. doi: 10.1093/molbev/msn083 PMID: 18397919.

47. Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87(1):23–9. PMID: 2110097.

48. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31(13):3406–15. PMID: 12824337; PubMed Central PMCID: PMCPMC169194.

49. Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009; 25(15):1974–5. doi: 10.1093/bioinformatics/btp250 PMID: 19398448; PubMed Central PMCID: PMCPMC2712331.